Math 343 / 643 Fall 2024
Final Examination Solutions

Professor Adam Kapelner

May 16, 2024

Full Name ____

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using an unauthorized cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold this Code of Academic Integrity.

______________________________ ______________

signature date

Instructions

This exam is 120 minutes and closed-book. You are allowed **three** pages (front and back) of a "cheat sheet", blank scrap paper (provided by the proctor) and a graphing calculator (which is not your smartphone). Please read the questions carefully. Within each problem, I recommend considering the questions that are easy first and then circling back to evaluate the harder ones. No food is allowed, only drinks.

Problem 1 Consider the independently realized Poisson with mean linear in x,

$$Y_i \overset{ind}{\sim} \text{Poisson}\left(\beta_0 + \beta_1 x_i\right) = \frac{(\beta_0 + \beta_1 x_i)^{y_i} e^{-(\beta_0+\beta_1 x_i) y_i}}{y_i!}$$

and assume flat priors on β_0 and β_1. We do not assume any structural equation model nor DAG for y and x. Let the units of x be centimeters and the units of y be kilograms.

(a) [3 pt / 3 pts] Demonstrate why a Gibbs sampler *cannot* be implemented to make inference for the parameter β_0.

$$\begin{aligned} f(\beta_0, \beta_1 \mid \boldsymbol{x}, \boldsymbol{y}) &\propto f(\boldsymbol{x}, \boldsymbol{y} \mid \beta_0, \beta_1) f(\beta_0, \beta_1) \propto f(\boldsymbol{x}, \boldsymbol{y} \mid \beta_0, \beta_1) \\ &= \prod_{i=1}^{n} \frac{(\beta_0 + \beta_1 x_i)^{y_i} e^{-(\beta_0+\beta_1 x_i) y_i}}{y_i!} \propto \prod_{i=1}^{n} (\beta_0 + \beta_1 x_i)^{y_i} e^{-(\beta_0+\beta_1 x_i) y_i} \\ f(\beta_0 \mid \beta_1, \boldsymbol{x}, \boldsymbol{y}) &\propto \prod_{i=1}^{n} (\beta_0 + \beta_1 x_i)^{y_i} e^{-\beta_0 y_i} \end{aligned}$$

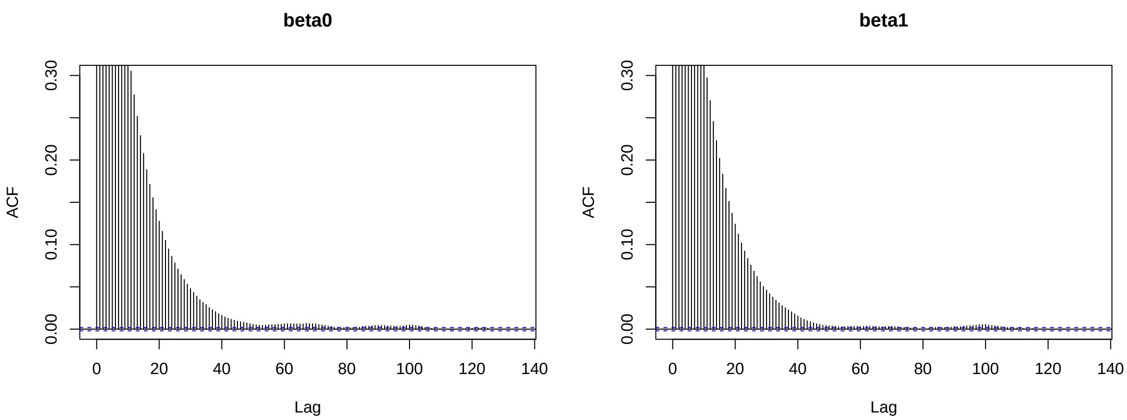
The above is not a kernel for any distribution we know of so we cannot use a Gibbs step.

Since we cannot use Gibbs sampling, we employ a Metropolis-Hastings sampler for the kernels of the conditional distributions of β_0 and β_1. Let $t \in \mathbb{N}$ indicates the iteration number of the sampler.

(b) [2 pt / 5 pts] Given the value of the previous iteration, $\beta_{0,t-1}$, propose a transition distribution by specifying its distribution and parameters.

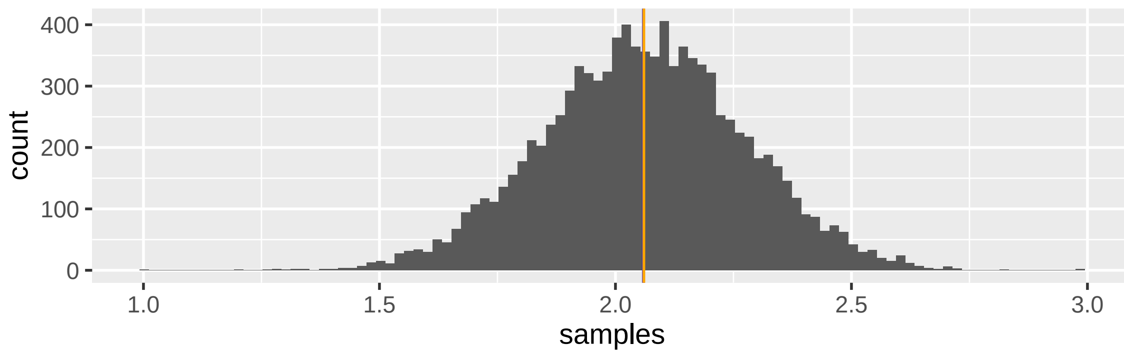
$\beta_{0,t} \sim \mathcal{N}\left(\beta_{0,t-1},\ 1\right)$

Below are the ACF plot for both parameters' samples:



(c) [1 pt / 6 pts] At what spacing should we thin the sample chains? 80

Below is the histograms of MCM samples for the parameter β_1 after properly burning and thinning the chain. The vertical line is the average of the chain's values.



(d) [5 pt / 11 pts] Write a detailed sentence that interprets the value of $\hat{\hat{\beta}}_1^{MMSE}$.

When comparing two observations A and B sampled in the same fashion as the observations in the historical dataset were sampled, when A has x 1cm larger than (B)'s x, then (A) is predicted to have an estimated mean count 2.05 ± 0.25 larger than (B)'s assuming the mean is linear in the p covariates.

Problem 2 There are many ways to measure to invest in the S&P500. Two popular tickers are SPY and VOO which have market caps of 500M and 1.1T respectively and have equally low expense ratios which are about 0.1%/yr. But are these two instruments equal? We pull the last ten years of data $n = 2581$ and we are interested in the response Y which is percent daily change. We choose to use the permutation test to test the difference. Let DGP_1 and DGP_2 denote the DGP's for SPY and VOO respectively. Let y_1 be the values for SPY and y_2 be the values for VOO.

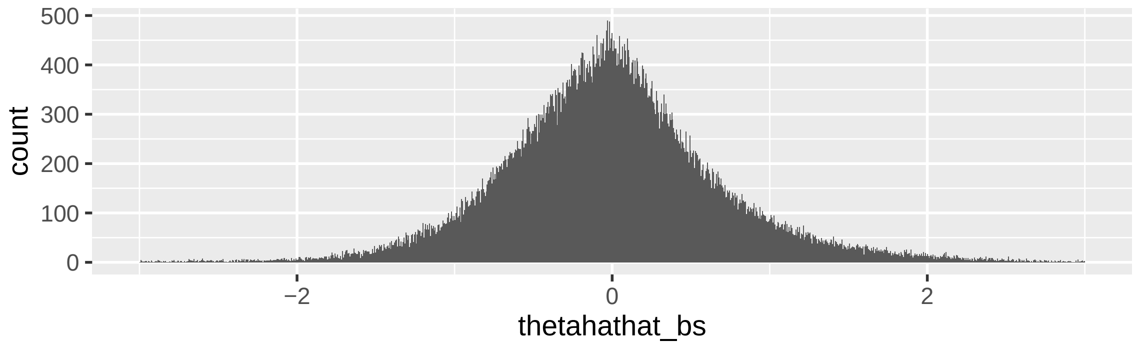
(a) [2 pt / 13 pts] What is the null hypothesis for this permutation test?

$H_0 : DGP_1 = DGP_2$

(b) [2 pt / 15 pts] During each iteration of the permutation test, how many numeric values are divided into two groups?

$2581 \times 2 = 5162$

We let $\bar{y}_1 - \bar{y}_2$ be the test statistic. The test statistic on the actual data is -0.00017. Over a total of $B = 100,000$ iterations, we have the following histogram of permutation test statistic values:



(c) [1 pt / 16 pts] Our choice of $B = 100,000$ is appropriate. Circle one: $\boxed{\text{yes}}$ / no

(d) [1 pt / 17 pts] The result of this test is... Circle one: $\boxed{H_0 \text{ retained}}$ / H_0 rejected

(e) [3 pt / 20 pts] Estimate a p-value for this test.

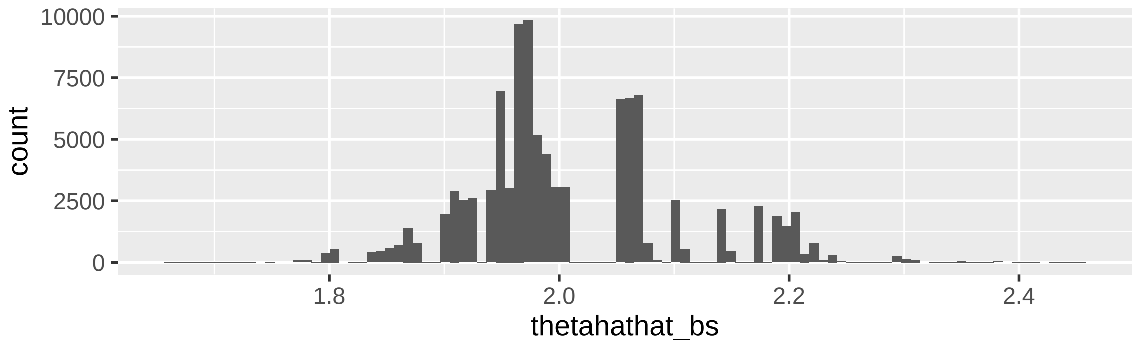
0.999

Now that we are reasonably convinced there's no difference between SPY and VOO, we turn to another question. We are interested in large quantiles of the percentage change. Let $\theta := \text{Quantile}[Y, 97.5\%]$. We wish to create a confidence interval for this parameter.

(f) [1 pt / 21 pts] Which statistical method / procedure provides asymptotically valid inference for θ? The answer should be one or two words only.

bootstrap

Assuming the correct answer to the previous question, we run this method and produce $B = 100,000$ iterations which we display below.



(g) [3 pt / 24 pts] Create an approximate 95% CI for θ.

$[1.85, 2.20]$

Problem 3 Consider $\boldsymbol{X}$ to be the the design matrix of for $n = 30$ observations and $p_{raw} = 5$ numeric covariates and their interactions. Let $\boldsymbol{H}$ be its orthogonal projection matrix. We assume also a continuous (real-valued) response model which is linear in these measurements,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}.$$

For the error term, we assume the "core assumption",

$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}_n\left(\boldsymbol{0}_n,\, \sigma^2 \boldsymbol{I}_n\right).$$

And the estimator for $\boldsymbol{\beta}$ is

$$\boldsymbol{B} := \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T\boldsymbol{Y}$$

(a) [4 pt / 28 pts] Find the distribution of $\boldsymbol{E}$, the vector of residuals. Show each step.

$$\begin{aligned} \boldsymbol{E} &= (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}) = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\mathcal{E}} = \\ (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\mathcal{E}} &\sim \mathcal{N}_n\left((\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{0}_n,\, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{I}_n(\boldsymbol{I}_n - \boldsymbol{H})^\top\right) = \mathcal{N}_n\left(\boldsymbol{0}_n,\, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})\right) \end{aligned}$$

We do not assume any structural equation model nor DAG for this phenomenon and observed measurements. We estimate $\boldsymbol{b}$ below along with selected inference information:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.656e+03  8.133e+03   1.187   0.2351
x_1         -1.034e+02  1.296e+02  -0.798   0.4251
x_2          2.322e+02  1.435e+02   1.619   0.1055
x_3          7.178e+03  3.477e+03   2.065   0.0390 *
x_4         -2.545e+04  3.573e+03  -7.123 1.07e-12 ***
x_5          1.983e+04  2.130e+03   9.310  < 2e-16 ***
x_1:x_2     -8.971e-01  2.306e+00  -0.389   0.6973
x_1:x_3      6.117e+00  4.022e+01   0.152   0.8791
x_1:x_4      3.227e+02  3.831e+01   8.423  < 2e-16 ***
x_1:x_5     -4.884e+02  2.005e+01 -24.356  < 2e-16 ***
x_2:x_3     -2.623e+02  2.771e+01  -9.465  < 2e-16 ***
x_2:x_4      8.294e+01  3.223e+01   2.573   0.0101 *
x_2:x_5      2.294e+02  3.071e+01   7.469 8.19e-14 ***
x_3:x_4      1.069e+03  2.968e+01  36.023  < 2e-16 ***
x_3:x_5      4.614e+02  3.787e+01  12.186  < 2e-16 ***
x_4:x_5     -8.674e+02  3.807e+01 -22.784  < 2e-16 ***

Residual standard error: 1450.0
Multiple R-squared:  0.87
```

Below are values of the 97.5%iles of the Student's T distribution (q) for many different degrees of freedom (df)

df	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
q	2.57	2.45	2.36	2.31	2.26	2.23	2.2	2.18	2.16	2.14	2.13	2.12	2.11	2.1	2.09	2.09

(b) [4 pt / 32 pts] Create a 95% CI for β the linear parameter for the interaction of $x_1 \times x_3$.

$$\begin{aligned} CI_{\beta,1-\alpha} &= \left[b \pm t_{n-(p+1),1-\alpha/2} s_b\right] = \left[b \pm t_{14,.975} s_b\right] = \left[6.117 \pm t_{14,.975} \cdot 40.22\right] \\ &= \left[6.117 \pm 2.14 \cdot 40.22\right] = \left[-79.95, 92.19\right] \end{aligned}$$

(c) [4 pt / 36 pts] [E.C.] Consider a new observation $\boldsymbol{x}_\star = [1\ 1\ 0\ 0\ 0]$. Create a 95% CI for $Y_\star$. Substitute all known quantities and use the notation in the problem header for all unknown quantitites.

(d) [2 pt / 38 pts] Circle one: $\boxed{R^2_{adj} < 0.87}$ / $R^2_{adj} = 0.87$ / $R^2_{adj} > 0.87$

(e) [3 pt / 41 pts] Compute the value of the $\hat{\hat{F}}$ statistic.

$$\hat{\hat{F}}^{-1} = \frac{p}{n-(p+1)}\left(\frac{1}{R^2} - 1\right) = \frac{15}{14}\left(\frac{1}{0.87} - 1\right) = 0.16 \quad \Rightarrow \quad \hat{\hat{F}} = 6.25$$

(f) [2 pt / 43 pts] Assume we now run the omnibus F test based on your computation in the previous question and we reject H_0. Also assume we did *not* make a Type I error. What can you now conclude about the vector $\boldsymbol{\beta}$? Make a numeric statement below. Hint: the answer is only a few characters. $\boldsymbol{\beta} \neq \boldsymbol{0}_{16}$

(g) [5 pt / 48 pts] The regression above shows $b_1 = -103.4$ and $s_{b_1} = 129.6$. Write the standard interpretation of b_1. Let the units of x_1 be centimeters (cm) and the units of y be kilograms (kg). Underline the words in this interpretation that *we know to be impossible given this specific regression.*

When comparing two observations A and B sampled in the same fashion as the observations in the historical dataset were sampled, when A has x_1 1cm larger than (B)'s x_1 value <u>and otherwise shares the same measurement values</u>, then (A) is predicted to have an estimated response 103.4 ± 129.6 kg lower than (B)'s assuming the mean is linear in the p covariates.

(The above is underlined since in a model that has interactions with the variable x_1, you cannot keep the value of $x_1 \times x_j$ constant for $j \neq 1$ when changing x_1).

Consider instead of using the estimator $\boldsymbol{B}$ above, we use the following estimator:

$$\boldsymbol{B}_{lasso} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{ (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}) + \lambda \sum_{j=1}^{p+1} |w_j| \right\} \quad \text{where} \quad \lambda > 0$$

(h) [5 pt / 53 pts] Using this new estimator, circle all the quantities below that are random:

$\boldsymbol{X}$ $\quad$ $\boxed{\boldsymbol{Y} \quad \boldsymbol{E} \quad \boldsymbol{\mathcal{E}} \quad \boldsymbol{B} \quad \boldsymbol{B}_{lasso}}$ $\quad$ $\boldsymbol{H}$ $\quad$ $\boldsymbol{\beta}$ $\quad$ σ^2 $\quad$ λ $\quad$ n $\quad$ p $\quad$ $\boxed{s_e \quad R^2}$

(i) [3 pt / 56 pts] Using $\boldsymbol{B}_{lasso}$, what is the most precise numerical statement you can say about r, the count of the number of rejections of $H_0 : \beta_j = 0$ where $j \geq 1$ at significance level $\alpha = 5\%$?

$r \leq 11$

Problem 4 Consider the lung dataset where missingness is dropped. Survival is measured in years. Below is the code to load the data and properly code it.

```
> lung = na.omit(survival::lung)
> lung$status = lung$status - 1 #needs to be 0=alive, 1=dead
> surv_obj = Surv(lung$time, lung$status)
```

This dataset came with measurements for each subject. We attempt to model survival using these features using the Weibull model employing log-linear link function we discussed in class. `Age` is measured in years and `meal.cal` is measured in cal/d. Below is the output with inference that employs the MLE core theorem:

```
survreg(formula = surv_obj ~ age + sex + meal.cal, data = lung)

                Value Std. Error     z      p
(Intercept)  6.16e+00   6.50e-01  9.48 <2e-16
age         -1.05e-02   8.24e-03 -1.28  0.202
sex=Female   3.44e-01   1.49e-01  2.30  0.021
meal.cal     8.54e-05   1.82e-04  0.47  0.639
Log(scale)  -3.00e-01   7.29e-02 -4.11  4e-05

Scale= 0.741
```

(a) [4 pt / 60 pts] Write an expression that estimates survival (in yr) for a 45yo male who eats 2000cal/d. Do not compute its value.

$$\hat{y} = e^{\boldsymbol{x}_* \boldsymbol{b}}\, \Gamma\left(1 + \frac{1}{\hat{\hat{k}}}\right) = e^{6.16 + 0.00105(45) + 0.0000854(2000)}\, \Gamma\left(1 + \frac{1}{0.741}\right)$$

(b) [4 pt / 64 pts] [E.C.] Evaluate if this Weibull model satisfies the proportional hazard assumption.

Below is the output from a cox proportional hazard model with inference that employs the MLE core theorem:

```
coxph(formula = surv_obj ~ age + sex + meal.cal, data = lung)

                  coef  exp(coef)  se(coef)       z  Pr(>|z|)
age          0.0160863  1.0162163  0.0111394  1.444    0.149
sex=Female  -0.4614061  0.6303966  0.1998968 -2.308    0.021 *
meal.cal    -0.0001175  0.9998825  0.0002485 -0.473    0.636
```

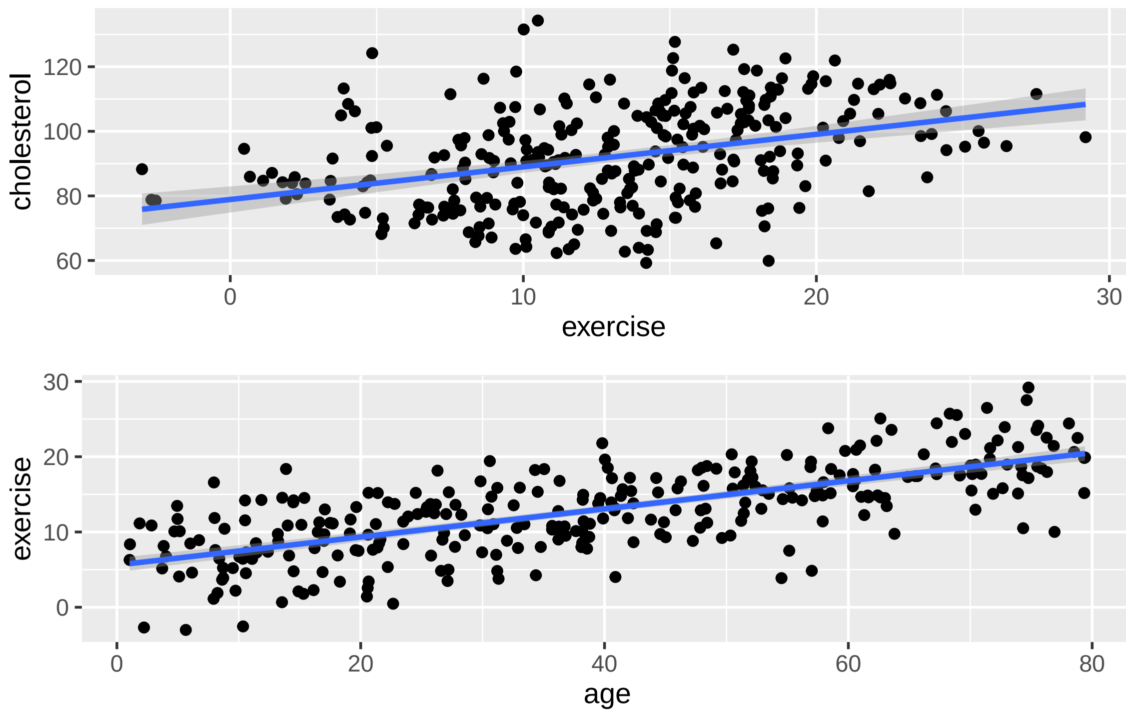
(c) [2 pt / 66 pts] Will the above model allow you to estimate of the survival (in yr) for a 45yo male who eats 2000cal/d? Circle one: yes / $\boxed{\text{no}}$

(d) [2 pt / 68 pts] Will the above model predict that the survival (in yr) for a 45yo male who eats 2200cal/d is shorter than the survival (in yr) for a 45yo male who eats 2000cal/d? Circle one: $\boxed{\text{yes}}$ / no

(e) [5 pt / 73 pts] Estimate how much more likely a 45yo female who eats 2000cal/d will survive the next week than a 45yo male who eats 2000cal/d (to the nearest two decimals).

$e^{-0.4614061} = 0.63$

Problem 5 We are interested in the affect of exercise on HDL cholesterol. We survey $n = 300$ people and measure their age (measured in years), exercise level (measured in average duration per day in minutes) and HDL cholesterol (measured in mg/dL). Below is a scatterplot of exercise on HDL cholesterol and a scatterplot of age on exercise:



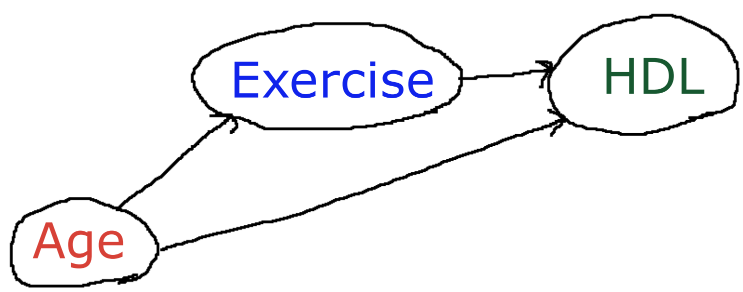
(a) [2 pt / 75 pts] Is there any way to prove absolutely that the correlation between exercise and cholesterol is spurious? Circle one: yes / $\boxed{\text{no}}$

(b) [1 pt / 76 pts] Is there any way to prove absolutely that the correlation between age and exercise is spurious? Circle one: yes / $\boxed{\text{no}}$

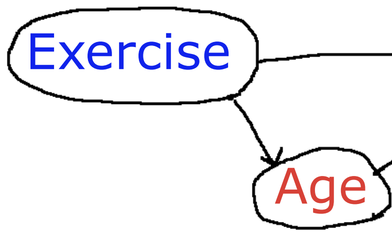
(c) [2 pt / 78 pts] Based only on the plots above and the situation described in the problem header, is there a way to definitively assess that the regression results of the first plot is a "Simpson's Paradox"? Circle one: yes / $\boxed{\text{no}}$

(d) [2 pt / 80 pts] Based only on the plots above and the situation described in the problem header, is there a way to definitively assess that the regression results of the first plot is a "Berkson's Paradox"? Circle one: $\boxed{\text{yes}}$ / no

(e) [4 pt / 84 pts] Draw below a DAG with nodes that include the variable names that could induce a "Simpson's Paradox" bias when investigating exercise as a cause of the phenomenon HDL.



(f) [4 pt / 88 pts] Draw below a DAG with nodes that include the variable names that could induce a partial blocking bias when investigating exercise as a cause of the phenomenon HDL.



Problem 6 We are interested in understanding the effect of a pill (coded per subject as $w_i = 1$) vs a placebo (coded per subject as $w_i = 0$) on lowering y HDL cholesterol (measured in mg/dL). We have $n = 100$ subjects. We assign subjects a w_i at the beginning of the study and we also record the subjects' sex, $x_i \in \{0, 1\}$, at the beginning of the study. There are 30 women and 70 men. Assume iid mean-centered noise and an additive treatment effect β_T which we called the PATE.

(a) [2 pt / 90 pts] Is this a controlled trial? Circle one: $\boxed{\text{yes}}$ / no

(b) [2 pt / 92 pts] Do we absolutely need to randomize the values of w_i to guarantee unbiased causal inference? Circle one: yes / $\boxed{\text{no}}$

(c) [2 pt / 94 pts] If we use "equal allocation", what is $\sum_{i=1}^{n} w_i = 1 \;\forall\; \boldsymbol{w}$? 50

(d) [2 pt / 96 pts] Assume we proceed with an equal allocation, completely randomized design. How many possible assignments are there? $\binom{100}{50}$

(e) [3 pt / 99 pts] Assume we proceed with an equal allocation, blocking design. How many possible assignments are there? $\binom{30}{15}\binom{70}{35}$

(f) [3 pt / 102 pts] Assume we proceed with an pariwise matching design. How many possible assignments are there? 2^{50}

(g) [4 pt / 106 pts] Why would rerandomization be a poor choice (when compared to blocking or pairwise matching) in this scenario?

Rerandomization will not guarantee x_1 averages are equal in the pill and placebo arms. Both blocking and pairwise matching will guarantee this.

(h) [2 pt / 108 pts] In order to test the sharp null, $H_0 : \forall i \; y_i(w_i = 1) = y_i(w_i = 0)$, which procedure can you use?

Fisher's Randomization Test